THE UNCONSTRAINED MINIMIZATION PROBLEM

Gary Charles Meyer

United States
Naval Postgraduate School

THESIS

THE UNCONSTRAINED MINIMIZATION PROBLEM

by

Gary Charles Meyer

Thesis Advisor: F. Faulkner

June 1971

Approved for public release; distribution unlimited.

The Unconstrained Minimization Problem

by

Gary Charles Meyer
Ensign, United States Navy
B.S., United States Naval Academy, 1970

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE
(with major in Mathematics)

from the

NAVAL POSTGRADUATE SCHOOL
June 1971

ABSTRACT

Considered within this paper is the problem of minimization of a function of unconstrained variables. A wide variety of solutions to this problem is presented and the possible advantages of each method are discussed. For the purpose of this paper these techniques are divided into four broad categories: general search directions; conjugate search directions; Cauchy's Steepest Descent and Newton's method; and variable metric methods.

TABLE OF CONTENTS

I. INTRODUCTION

The problem under consideration in this paper is that of minimizing f, a function of n unconstrained variables. This is a problem that arises frequently in many widely varied fields. In general, it may be more likely to find that the variables of the function have certain constraints placed upon them. While this problem is conceptually more involved it is felt that the key to its solution lies in the solution of the unconstrained problem. Thus a great deal of attention has been given to this latter problem. It is felt that if the unconstrained problem can be solved then its method of solution can be applied to the constrained problem, by the technique of adjoining penalty functions corresponding to the constraints.

The minimization of a function is certainly not a new problem to mathematics. Famous scientists such as Cauchy and Newton long ago devised methods of solution. In fact, their methods remain useful today and will be two of those discussed in this paper. However, the functions to be minimized have become more and more complex. The classical methods of Cauchy and Newton are often found to be inadequate.

Thus since the late 1950's there has been a great deal of research in this field. The key to this surge has been the computer. Difficult and complex methods of solution would be useless without the computer; but its availability

has allowed the development of many new ingenious minimization techniques which would have previously been impossible to implement.

With this new research, one fact has become apparent. No one method seems to be the most efficient for every type of function. Most methods have certain characteristics which make them more useful in some cases than others. Therefore, apparently there is no simple solution to the minimization problem.

Thus, the purpose of this paper is to present a wide selection of these new and old methods of solution. Their relative advantages and disadvantages will be discussed in the hope that the most efficient method can be selected for the function to be minimized. Unfortunately the complexity of this field and the lack of sufficient time has not permitted the author to program many of the methods to be discussed. Since this is important to the rating of the relative efficiencies of these methods, the findings of other researchers will be used and referenced to allow further in depth study of specific problems.

II. GENERAL SEARCH DIRECTIONS

A. GRID METHOD [9]

As we go from the first to the fourth category we go to more and more sophisticated methods. The first method is quite simple, but its applicability is very limited. Its advantage is the ease with which it can be programmed.

This method, to be useful, requires that there be some knowledge of the location of the minimum. Therefore, assume that the minimum, $X = (x_1,\ldots,x_n)$, is known to lie within a certain rectangular region defined as follows:

$$a_i \leq x_i \leq b_i \qquad i = 1,\ldots,n ;$$

in which a_i and b_i are known. Define a grid over the region as follows. Let

$$d_i = b_i - a_i \qquad i = 1,\ldots,n$$

$$s_i = d_i/r_i \qquad i = 1,\ldots,n ;$$

where r_i is a positive integer.

The function is evaluated at each point of this grid and the smallest value is taken as the minimum of the function. The difficulties with this method are obvious. To obtain a good approximation to the minimum it would be necessary to make each s_i "small." But decreasing the size of s_i increases the number of points at which the function must be evaluated. The number of evaluations needed is $M = (r_1+1)(r_2+1)\ldots(r_n+1)$. If n is large then

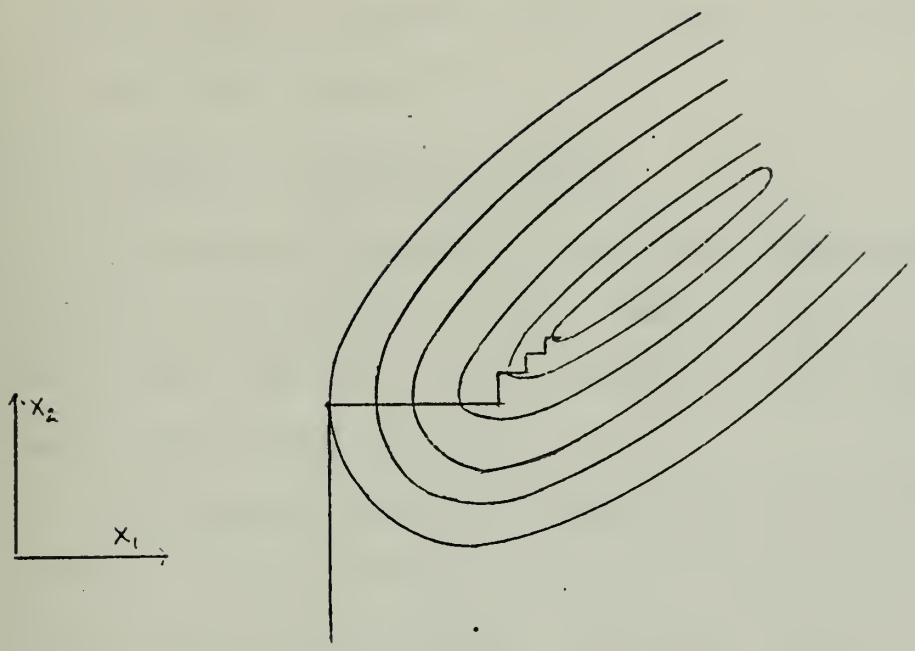
M can be so great that the method would require far too many evaluations for it to be useful.

B. ALTERNATING VARIABLE METHOD

In this method a basic technique is introduced that will be employed in most of the methods that follow. A point and a direction are chosen in some way, and the minimum value is sought along the resulting line. A direction is a vector in n-space which is usually represented by a column matrix. At this time it will be assumed that, given a straight line, the point on that line at which the function has a minimum value can be determined. Since this is a secondary problem and its solution is rather elementary it will be dealt with later, in Appendix A. Included there are some techniques for treating the basic problem.

This minimization process is characterized by the use of permanently fixed search directions. Generally, the directions are chosen parallel to the various co-ordinate axes. Each variable in turn is changed or perturbed and a search carried out so as to minimize the function on that line. The effect, as shown below, is that of a staircase, in which the steps decrease in size near the minimum, if the function is a quadratic.

In n dimensions, a function whose level surfaces were hyperspheres would be minimized in n searches. But if a function is not of this nice nature then certain difficult



problems can arise. For practical use there must be some means of determining when the process should be halted. There are many such rules, called convergence criteria, which can be used. One of these criteria is based upon the change in the value of the function over one iteration [9]. It has been suggested for some methods that if this change is less than some ε then the process should be terminated. But one of the problems that may be encountered in the Alternating Variable Method is that if the principle axis of the function is not aligned, at least approximately with one of the co-ordinate axes, progress along each search direction may be very small. In this case the choice of the convergence criterion discussed above may lead to a halt of the process well before the actual minimum is reached. Since many functions arising in problems are not of the "nice" hyperspherical nature,

the Alternating Variable Search Method is often too simple for good results [9].

C. METHOD OF HOOKE AND JEEVES

The method invented by Hooke and Jeeves attempts to improve the inflexible search routine discussed above. To do this the concepts of exploratory moves and pattern moves are introduced [3,9].

The exploratory process resembles the alternating variable search technique in that it uses the co-ordinate directions to search along. However, it is not assumed that the minimum along each line can be found. Instead x_i is perturbed by an amount d_i while the other variables are held fixed. If the functional value is decreased with this step then the new point replaces the previous one and the next variable is considered. If the function is not decreased then the original x_i is perturbed by $-d_i$, and again the functional values are compared. This new point may or may not replace the old one, but in either case the next variable is then considered. One cycle is complete when all the variables in turn have been perturbed.

The next step taken is what Hooke and Jeeves call a "pattern" move, which is made from the last point arrived at during the exploratory phase. Let us call this last point a_n and let a_0 be the point at which the cycle started. The pattern move will then be to the point $2a_n - a_0$ The purpose

of this is to make another move in the general direction of total progress made during the previous cycle. From this point a new round of exploratory moves is performed and the functional value at the last point is compared with the value of the function at a_n. The entire process is then repeated from the point which had the smaller functional value.

This is continued until no progress is made during a cycle of exploratory moves, which may indicate that the present point is within d_i of the minimum or it may be that the minimum point lies in a steep skew valley. For further progress then, d_i must be reduced before the process is continued. When d_i becomes less then some specified ε it is assumed that the operation has converged.

A slow rate of convergence is often a very real problem with this method. The choice of d_i is critical. If the initial point is far from the minimum and d_i is relatively small then the process would be very time consuming with a great number of functional evaluations needed.

Even with its disadvantages, this method is the first example of a principle which will be applied over and over later in this paper. The method attempts to use past information that has been obtained about the function. Thus the process calls for a move in the direction of progress during the exploratory moves which were made to indicate the general local nature of the function. The incorporation of any previous knowledge obtained about the function

is an important characteristic of most of the more involved minimization methods.

D. SIMPLEX METHOD: NELDER AND MEADE (1965)

In this method a simplex is defined: this is initially a configuration of n+1 equally spaced points in n space; for example an equilateral triangle in two space.

The method presented by Nelder and Meade was designed to eliminate some of the problems that arose in early simplex methods [3,9]. Unlike some of these other methods this one does not require that the simplex remain equilateral. With this greater flexibility in shape it may possibly be easier for the method to follow the contours of the function and thus not be obstructed by something such as a steep skew valley.

The first step is to evaluate the function at each of the vertices of the simplex. The vertex at which the function is maximum, V_1, is then reflected through the centroid, C, of the other vertices.

$$V_{new} = (1+\alpha)C - \alpha V_1$$

or

$$\alpha = \frac{|V_{new} - C|}{|C - V_1|} .$$

If $\alpha=1$ this method is simply one of the earlier methods in which the simplex remains equilateral.

After the function is evaluated at V_{new} this value is compared with the values at the other vertices. There are four different cases which must be considered.

i. First assume that $f(V_{new})$ is less than the previous second largest value of the function at a vertex but larger than the value at some other vertex. Since the point at which the function was largest was the point that was reflected this case implies that the second largest value of the function has become the largest. Thus V_{new} replaces V_1 and the process continues.

ii. Second, assume $f(V_{new})$ is less than the value of the function at all the vertices. This indicates that a direction has been found along which the function can be greatly reduced. Thus it might be wise to investigate this direction, further, which would not be the case if condition i. held. To take a step further in this direction therefore calls for the use of an expansion coeffi-cient $\gamma > 1$.

Define:

$$V_e = \gamma V_{new} + (1-\gamma)C \ .$$

The result of this process is to define a new point V_e which is on the same line with V_{new} but is farther from the centroid. If the value of the function at V_e is less than $f(V_{new})$ then V_e replaces V_1. Otherwise the point V_{new} replaces V_1. In either case the process is then continued.

iii. Third, assume $f(V_{new}) < f(V_1)$ but $f(V_{new})$ is still greater than the values of the function at the other vertices. This could indicate that there is a relative

minimum somewhere between V_{new} and V_1. Thus it may be desirable not to go quite as far as called for by the reflection. This can be accomplished through the use of a contraction coefficient $\beta < 1$.

Let us take:

$$V_c = \beta V_{new} + (1-\beta)C, \text{ where } 0 < \beta < 1.$$

Again the function is evaluated at this new point and this value is compared with the values of the function at the other vertices. If the result is still a maximum then the contraction is considered a failure and a different strategy must be used. Otherwise V_c replaces V_1 and the process continues.

A failure in the contraction could indicate that the simplex has entered a steep skew valley or that a minimum is being approached. In either case the size of the simplex must be reduced so that further progress may be made. The natural way of doing this is to cut by one half the distance of each vertex from the vertex at which the function was a minimum. Thus the simplex is shifted toward what should be a more favorable area. After doing this the reflection process is again continued.

iv. Fourth, assume $f(V_{new}) > f(V_1)$. In this case the reflected direction does not seem very favorable so it is rejected. Again a reduction in the size of the simplex is called for before the procedure is continued.

The following two criteria could be used to halt the process:

i. Let

$$s = \sum_{i=1}^{n+1} \frac{(f_i - \bar{f})^2}{n}$$

be the standard deviation. Then if s is less than some specified number, convergence might be assumed. In a steep skew valley this criterion could cause a premature halt and thus if this problem is feared the following criterion should be employed.

ii. Calculate s after each k function evaluations. Convergence would be assumed if successive s's were less than some specified number and the difference between two successive $\bar{f}$'s was less than some small number.

This simplex method is best suited for problems in which the number of variables is small. The process is relatively slow and with a large number of variables the required computer time could become intolerably large. One technique that should probably be checked is whether it might not be better to reflect through the centroid of some of the vertices with smaller function values rather than the centroid of all the vertices. Also in the expansion phase of the process it would probably be better to continue expanding until a failure is achieved. Since this direction is favorable why should only one expansion be attempted? Since this expansion would result in a new point that may be quite a distance from the remainder of the simplex it would then be wise to reduce the

distances of the other vertices from this point. But then this has the effect of shifting the simplex toward what should be a more favorable area.

E. ROSENBROCK'S METHOD

This method, devised by Rosenbrock, is a rather obvious development from the method of Hooke and Jeeves discussed earlier [3, 14]. The process is usually started by using the co-ordinate directions as the first search directions but, in general, any set of n mutually orthonormal direction vectors could be used. As with the method of Hooke and Jeeves, each direction is considered individually with a step of length d_i taken along it. If the value of the function at this new point is less than or equal to the value at the original point then the step is termed a success. Otherwise it is considered a failure. If a success had resulted then d_i is multiplied by some $\alpha > 1$. If the result was a failure then d_i is multiplied by $-\beta$, $0 < \beta \leq 1$. In either case the next search direction is then investigated. This procedure is continued until a success and a failure have been obtained in each direction. This constitutes the end of one stage.

After each stage is completed new search directions must be defined. This is done through the use of the following vectors:

$$a_i = \sum_{k=i}^{n} \Delta_k \, \xi_k^j, \quad i = 1,\ldots,n \; ;$$

where ξ_k^j, a unit vector, is the k'th search direction in the j'th stage; Δ_k is the sum of all the steps taken in the direction of ξ_k^j. From this definition it is readily apparent that a_1 is the total progress made during that stage; a_2 is the total progress made in all the directions other than the first, etc. An important property of these vectors is that they are linearly independent. This property results from the choice of the definition for a success. The linear independence property of the a_i's would be lost if at any time there was no progress made along one of the search directions during a stage. At first this seems to be a possibility since it could happen that a stage is started at a point that minimizes the function along one of the search directions. But allowing equality in the definition of a success eliminates this problem. During the process the step size along such a direction would be reduced to such an extent that, for computer use, the value of the function at these two points would be the same and thus a success is obtained. While this step may get very, very small it will still be different from zero and thus some progress is always made in every direction.

The new directions are defined as follows by the Gram Schmidt orthogonalization process.

$$s_k = a_k - \sum_{\ell=1}^{k-1} \left(a_k' \, \xi_\ell^j\right) \xi_\ell^j$$

and

$$\xi_k^{j+1} = s_k / ||s_k||.$$

Thus these new vectors form a set of n mutually orthonormal search directions.

There are various criteria that could be used to stop this process. A limit could be set upon the number of function evaluations to be made during the process. This obviously may halt the method well before the minimum is reached, but it is helpful in avoiding the use of too much computer time. The process could also be terminated if $||a_1||$ is smaller than some given number [3]. This would mean that the progress made during one stage was very small. This seems to be a quite natural stopping criterion, for surely as the minimum is approached the progress made will get less and less. Unfortunately this could also be the characteristic for a steep skew valley. If there is a possibility that the function has this property then great care must be taken to avoid a premature halt in the process. A third criterion for convergence could be $|a_2|/|a_1| > .3$. This should be used only if the d_i's are scaled to have similar magnitudes [3]. The reasoning behind this criterion is that $|a_2|/|a_1| > .3$ indicates that the direction of total progress is rapidly changing which is again a characteristic trait of the function in the vicinity of the minimum. This rapid change could also be present early in the process so this convergence criterion should be applied only after a number of stages have been completed.

There is a close relationship between the pattern move devised by Hooke and Jeeves and the ξ_1^j defined above. They

are both in the direction of total progress made during one stage. Rosenbrock's method is far superior, though, because of its complete use of the knowledge gained about the function as is evident in the generation of the new search directions [3,9]. This has the property of aligning the search directions with the principle axis of the function.

The ease with which this method can be adapted to computer use and its relative stability has shown it to be one of the most useful of the direct search procedures.

F. DAVIES, SWANN AND CAMPEY (1964)

This method is a further refinement of the useful method invented by Rosenbrock [3,9]. It attempts to remove the restriction of a fixed step length. As with Rosenbrock's method, the search directions are n mutually orthonormal vectors chosen initially as parallel to the co-ordinate axes. In this case, though, it will be assumed that, within a certain degree of accuracy, the minimum along each search direction can be found. This assumption introduces one difficulty that Rosenbrock's method does not have.

It may not be possible to make progress along a certain search direction and thus the vectors, a_i, defined in Rosenbrock's method, would not be linearly independent. Assume that there can be formed (n-m) linearly independent vectors. These vectors are orthogonalized as described previously and become (n-m) search directions. The other

m directions needed are those along which no progress was made during the previous cycle. Since none of the new vectors have components in the direction of these m vectors and the m vectors were already orthonormal it is evident that again the process is begun with n mutually orthonormal vectors. Convergence criteria similar to those suggested with Rosenbrock's method can be applied to this method.

The assumption that was made in regard to minimization along a line introduces an added factor that must be considered when choosing between this method and the one devised by Rosenbrock. Depending upon the technique used to obtain this minimum a larger number of function evaluations might be needed. Thus, if this latest method did not significantly increase the rate of convergence the additional evaluations required might indicate that Rosenbrock's method is better in that case. Also when far from the minimum there is no real advantage to obtaining the exact minimum along any specific direction. Thus again the fixed step length may have an advantage because of the less time required. In general, though, this method has been found superior to the method of Rosenbrock [3,9].

G. MATRIX ESTIMATOR

In this section it will be assumed implicitly that the function may be approximated by a quadratic, at least in some region. A method is developed for determining the

coefficients and using these to estimate the point where f effects its minimum [6].

Let us now consider the case in which the function f is a quadratic in the form:

$$f(x) = \tfrac{1}{2}x'Ax + b'x + c$$

where $x' = (x_1,\ldots,x_n)$.

Under these conditions if the matrix A is known then it may be shown that the minimization problem is easy to solve, as follows. Let $\bar{x}$ be the point where the minimum occurs and consider:

$$\nabla f(x) = b + Ax$$

and

$$\nabla f(\bar{x}) = 0 = b + A\bar{x}.$$

If these are subtracted we get:

$$\nabla f(x) - \nabla f(\bar{x}) = Ax - A\bar{x}$$

whence

$$\bar{x} = x - A^{-1}\nabla f(x). \qquad (1\text{-}1)$$

Thus (1-1) can be applied to find $\bar{x}$ if A and $\nabla f(x)$ can be determined. The matrix A can be found as follows:

Consider a sequence of points $x^{k+1} = x^k + p_k$ where $p_k = \lambda^k d_k$ and λ^k is selected to minimize f along the line defined by x_k and d_k. Then,

$$f(x^{k+1}) = c + b'(x^k+p_k) + \tfrac{1}{2}(x^k+p_k)'A(x^k+p_k)$$

$$= f(x^k) + (b'p_k+p_k'Ax^k) + \tfrac{1}{2}p_k'Ap_k$$

$$= f(x^k) + \nabla'f^k p_k + p_k'Ap_k/2 \ . \qquad (1\text{-}2)$$

$$f(x^k) = c + b'(x^{k+1}-p_k) + \tfrac{1}{2}(x^{k-1}-p_x)'A(x^{k+1}-p_k)$$

$$= f(x^{k+1}) - \nabla'f^{k+1}p_k + \tfrac{1}{2}p_k'Ap_k/2 \qquad (1\text{-}3)$$

but $\nabla'f^{k+1}p_k = 0$ because f was minimized along d_k.
Hence,

$$f(x^{k+1}) = f(x^k) - p_k'Ap_k/2. \qquad (1\text{-}4)$$

This last relation is very useful in determining the elements of the matrix A, as follows.

Let us define e^k to be a column vector that is zero except for a one in the k^{th} position and choose $d_{k-1} = e^k$. For this choice of d_k equation (1-4) reduces to:

$$\frac{-2(f(x^{k+1}) - f(x^k))}{(\lambda^k)^2} = a_{k+1,k+1} \quad k = 0,1,\ldots,n-1. \qquad (1\text{-}5)$$

Where a_{ii}, $i = 1,\ldots,n$ are the diagonal elements of A. The above equation thus generates the diagonal elements of A by the use of function values only.

To obtain the off-diagonal elements let us define:

$$d_{ij} = e^i + e^j \text{ for } i = 1,\ldots,n-1;\ j = i+1,\ldots,n.$$

Then let λ^{ij} be the scalar that minimizes f along d_{ij}. This reduces equation (1-4) to the form:

$$\frac{(-f(x^{k+1}) + f(x^k) - (\lambda^{ij})^2 a_{jj} - (\lambda^{ij})^2 a_{ii}}{2(\lambda^{ij})^2} = a_{ij} \qquad (1\text{-}6)$$

$i = 1,\ldots,n-1$

and

$j = i+1,\ldots,n.$

Equations (1-5) and (1-6) thus completely define the matrix A. This procedure is accomplished by minimization along $n(n+1)/2$ directions. If it happens that one of the λ's is zero then it is assigned the value of a very small but nonzero constant and equations (1-5) and (1-6) are used to generate approximations to a_{ii} and a_{ij}. After A has been found the problem remaining is the determination of $\nabla f(x)$.

For the choice of $d_{k-1} = e^k$:

$$\frac{\partial f(x^n)}{\partial x_i} = \frac{\partial f(x^i)}{\partial x_i} + \sum_{j=1}^{n} a_{ij}(x_j^n - x_j^i) \qquad (1\text{-}7)$$

if f is quadratic.

But $\partial f(x^i)/\partial x_i = 0$; $x_j^n = x_j^i$ for $j = 1,\ldots,i$; and $x_j^i = x_j^0$ for $j = i+1,\ldots,n$ by our choice of d_k. The above conditions reduces (1-7) to the following:

$$\frac{\partial f(x^n)}{\partial x_i} = \sum_{j=i+1}^{n} a_{ij}(x_j^n - x_j^0). \qquad (1\text{-}8)$$

Complete knowledge of A thus enables the calculation of the gradient at any point in the above sequence. A development very similar to that used to derive (1-1) produces the following results:

$$\nabla f(y) = \nabla f(x^n) + A(y-x^n) \tag{1-9}$$

$$\bar{x} = y - A^{-1}\nabla f(y) \tag{1-10}$$

In general, of course, most functions of interest are not quadratic. But since in the neighborhood of the minimum of the function we will assume generally they closely approximate a quadratic, it may be possible to adapt the above procedure to an iterative process. Each $n(n+1)/2$ searches would produce a new approximation to A. By using this approximation in equations (1-9) and (1-10) it may be possible to approach the minimum. Unfortunately there is no guarantee that far from the minimum this method would produce a good or even useful approximation to the matrix A. Therefore if only function values are to be used it would probably be best to employ one of the other methods, such as Rosenbrock's, until it is felt that the process has reached a point that is reasonably close to the minimum. Switching over to this latter method at this time could possibly be very valuable because of its exact nature for quadratic functions. Of course, in the use of this method it must be realized that more storage space will be required of the computer, since the matrix A must be stored. And since each cycle of this method requires $n(n+1)/2$ searches rather than the previously used n directions, significant progress must be made at each stage to warrant its use.

III. CONJUGATE SEARCH DIRECTIONS

In this chapter we will consider some methods based on the idea of conjugate directions.

Let us again consider a quadratic function of the form:

$$f(x) = c + b'x + x'Ax/2.$$

Two directions d_i and d_j are called conjugate with respect to A if

$$d_i A d_j' = 0, \text{ for } i \neq j.$$

Conjugate directions play a significant role in recent developments in minimization theory, as shall be seen in the following discussion. Assume that $d_1, \ldots, d_n$ are n mutually conjugate directions. Let $\bar{x} = x^0 + \sum_{i=1}^{n} \lambda^i d_i$ where the λ^i's are selected to minimize $f(x^0 + \sum_{i=1}^{n} \lambda^i d_i)$. We shall see that the resulting point furnishes the desired minimum to f.

Consider:

$$f(x) = \tfrac{1}{2}(x^0 + \sum_{i=1}^{n} \lambda^i d_i)'A(x^0 + \sum_{i=1}^{n} \lambda^i d_i) + b'(x^0 + \sum_{i=1}^{n} \lambda^i d_i) + C$$

$$= f(x^0) + \sum_{i=1}^{n} (\tfrac{1}{2}\lambda_i^2 d_i' A d_i + \lambda_i d_i'(Ax^0 + b)).$$

Therefore to select the λ^i's to minimize $f(x + \sum_{i=1}^{n} \lambda^i d_i)$ is the same problem as selecting each λ^i to minimize $(\tfrac{1}{2}(\lambda^i)^2 d_i' A d_i + \lambda^i d_i'(Ax^0 + b))$. Therefore the choice of each λ^i is independent of every other λ^j, $j \neq i$.

This implies that if n conjugate directions are used it is sufficient to minimize once along each direction to obtain the minimum value of the function. Since any arbitrarily chosen n linearly independent vectors usually do not have this property the advantage in using conjugate directions is clearly evident. Let us now take up methods which in one way or another make use of searches in conjugate directions.

A. POWELL'S METHOD

This method depends upon the following manner of generating conjugate directions. Assume that the function is a positive definite quadratic. Let us pick a direction d_1, and two points x^0 and x^2 such that x^0-x^2 is not a multiple of d_1. Let us define:

$$x^1 = x^0 + \alpha d_1$$

$$x^3 = x^2 + \beta d_1$$

where α and β are chosen such that $f(x^0+\alpha d_1)$ and $f(x^1+\beta d_1)$ are the minimum values on their respective lines.

Then

$$d_1' \nabla f(x^3) = d_1'(Ax^3+b) = 0$$

and

$$d_1' \nabla f(x') = d_1'(Ax'+b) = 0$$

so that

$$d_1' A(x^3-x') = 0. \qquad (2\text{-}1)$$

Equation (2-1) shows that the direction (x^3-x^1) is conjugate to the original search direction. It is this reasoning that Powell used to develop his method [13].

As with most methods involving search directions this procedure is begun with a choice of n linearly independent search directions $d_1^1,\ldots,d_n^1$. The function is then minimized along each direction in succession, with x^n being the point resulting from the minimization along d_n^1, the last search direction. From this the direction (x^n-x^0) is obtained where x^0 is the initial point. This direction is used as another search direction along which to minimize. The result of this is a new initial point from which to begin another round of searches along n linearly independent directions. The last n-1 directions are retained but advanced in index by one as follows:

$$d_j^2 = d_{j+1}^1 \quad j = 1,\ldots,n-1$$

$$d_n^2 = x^n-x^0.$$

There is a possibility that no progress might be made along a certain search direction during any given cycle. For example, assume that a cycle is begun at a point which minimizes the function along d_1^i. Therefore no further progress is made along this direction which means that (x^n-x^0) will have no component along d_1^i. But then d_1^i is deleted from the next round of searches which implies that the n search directions will not span the given space. If this

problem does not arise then the method will generate n conjugate directions after n stages. The next search stage would therefore produce the desired minimum regardless of where the initial point was located. Of course, all this was done under the assumption that the function was a positive definite quadratic. Since, in general, this is not in fact the case then the process must be applied iteratively.

The problem of loss of linear independence is a serious one, though, and cannot be overlooked. If this problem arose the method could not converge no matter how long the computer worked on it. Since conjugate directions seem to offer significant advantages in the minimization problem a refinement of the above method was sought to eliminate its flaws. Just such a method was suggested by Powell in 1964.

B. REVISED CONJUGATE DIRECTIONS BY POWELL [12]

Again it is assumed that the method is begun with n search directions $d_1^1,\ldots,d_n^1$; these and each subsequent set are to be linearly independent and scaled such that:

$$d_j^{i\prime} A d_j^i = 1 \quad \text{for } j = 1,\ldots,n.$$

Let $\det D = \det(d_1^i \cdots d_n^i)$. It will now be shown that this determinant is maximized when the d_j^i's are mutually conjugate. Let v^i, $i = 1,\ldots,n$, be a set of n conjugate, nonzero scaled vectors. Since they are conjugate and nonzero they

must be linearly independent. This implies that each d_j^i can be written as a linear combination as follows:

$$d_j^i = \sum_{k=1}^{n} u_{jk} v^k$$

or

$$(d^i \ldots d_n^i) = (\sum_{k=1}^{n} u_{1k} v^k \ldots \sum_{k=1}^{n} u_{nk} v^k).$$

Therefore

$$\det D = |(v^1 \ldots v^n)||U| \tag{2-2}$$

$$d_j^i{}' A d_k^i = (\sum_{m=1}^{n} u_{jm} v^m)' A (\sum_{p=1}^{n} u_{kp} v^p)$$

$$= \sum_{m=1}^{n} \sum_{p=1}^{n} u_{jm} u_{kp} v^{m'} A v^p$$

$$= \sum_{m=1}^{n} u_{jm} u_{km} v^{m'} A v^m$$

since

$$v^{m'} A v^p = 0 \text{ for } m \neq p.$$

Therefore

$$d_j^i{}' A d_j^i = 1 = \sum_{k=1}^{n} u_{jk} u_{jk}. \tag{2-3}$$

But equation (2-3) shows that the determinant of U can not exceed one, which it equals only if U is an orthogonal matrix. If this is the case:

$$d_j^i{}' A d_k^i = \sum_{m=1}^{n} u_{jm} u_{km} = 0 \quad j \neq k. \tag{2-4}$$

Equation (2-4) implies therefore that the directions $d_1^i, \ldots, d_n^i$ are mutually conjugate. Therefore since the v^i's were chosen arbitrarily it can be seen from (2-2) that det

D is maximized when the determinant of U is maximized, which implied that the given directions were conjugate. This result then forms the basis for this new method devised by Powell.

For the most part this new method resembles the first method by Powell: after the n search directions are used, it is desired to look at the direction x^n-x^0. Again this direction is examined since it appears to be along this direction that the process is progressing toward the minimum. Unlike the earlier method, though, this new direction is not automatically accepted as a new search vector. It must first be determined whether replacing one of the vectors in D by this new vector would increase det D. If it is increased then it can be reasoned that the directions must be approaching conjugacy. Obviously the det D would not increase if the replacement made det D = 0. This would be the case if the new direction was not linearly independent. Thus by using the new direction only when it increases det D insures that the linear dependence problem of the earlier method is eliminated. The question then arises as to which of the old directions should be deleted when the new direction is added.

If we assume that the vectors are scaled such that $d_j^{k\prime}Ad_j^k = 1$ then

$$f(x^{i-1}) = \tfrac{1}{2}(x^i-\lambda^i d_i^k)'A(x^i-\lambda^i d_i^k) + b'(x^i-\lambda^i d_i^k) + C$$

$$f(x^i) = \tfrac{1}{2}(x^i)'Ax^i + b'x_i + C$$

$$f(x^{i-1}) - f(x^i) = \tfrac{1}{2}(\lambda^i)^2 d_i^{k\prime} A d_i^k - (\lambda^i d_i^{k\prime} A x^i + \lambda^i d_i^{k\prime} b)$$

$$= \tfrac{1}{2}(\lambda^i)^2 - \nabla' f(x^i) d_i^k = (\lambda^i)^2 .$$

Therefore

$$\lambda^i = \sqrt{2(f(x^{i-1}) - f(x^i))}. \qquad (2\text{-}5)$$

Now $(x^n - x^0) = \lambda^1 d_1^k + \ldots + \lambda^n d_n^k = \mu d_p^k$, where μ is chosen so that $d_p^{k\prime} A d_p^k = 1$.

If d_p^k replaces d_i^k in det D the following results:

$$\left| d_1^k \ldots d_{i-1}^k d_p^k d_{i+1}^k \ldots d_n^k \right| = \left| d_1^k \ldots d_{i-1} \left(\frac{\lambda^1}{\mu} d_1^k + \ldots + \frac{\lambda^n}{\mu} d_n^k \right) \ldots d_n^k \right|$$

$$= \left| d_1^k \ldots d_{i-1}^k \frac{\lambda^i}{\mu} d_i^k d_{i+1}^k \ldots d_n^k \right|$$

$$= \left| \frac{\lambda^i}{\mu} \right| (\det D).$$

From this it can be seen that replacing d_i^k by d_p^k has the effect of multiplying the determinant by $|\lambda^i/\mu|$. This multiplication factor is greatest when the largest λ^i is chosen. But, since λ^i represents the change in the value of the function when minimizing along d_i^k, the largest λ^i corresponds to the direction along which the function underwent the greatest reduction in value. Thus the new direction should replace whichever direction produced the greatest reduction in the value of the function as long as $|\lambda^i/\mu| \geq 1$. If this last inequality is not satisfied for some i, this substitution would in all instances reduce the value of det D which

is contrary to the desired results. As with Powell's first method, this process also calls for minimization along $x^n - x^0$ starting from x^0 to obtain a new point from which to begin the next round of searches. This step, though, involves further problems that must be considered.

Using the three points x^0, x^n, and $2x^n - x^0$ let us use a quadratic interpolation to obtain the minimum along the line joining x^n and x^0. Since these three points are equally spaced the following function can be used for this purpose:

$$g(t) = at^2 + bt + c \quad \text{for } 0 \leq t < \infty$$

where

$$g(-1) = g_1 = f(x^0)$$

$$g(0) = g_2 = f(x^n)$$

and

$$g(1) = g_3 = f(2x^n - x^0).$$

Solving a system of three equations in three unknowns produces the following results:

$$a = (g_1 - 2g_2 + g_3)/2$$

$$b = (g_3 - g_1)/2$$

$$c = g_2.$$

The value of t, t_s, for which g(t) has its minimum(maximum) value can be found by setting $\frac{d}{dt}(g(t))$ equal to zero. Hence,

$$t_s = -b/2a = (g_1 - g_3)/(2(g_1 - 2g_2 + g_3)).$$

The value of g at this point is:

$$g_s = at_s^2 + bt_s + c$$

$$= g_2 - (g_1 - g_3)^2/(8(g_1 - 2g_2 + g_3)).$$

It must first be insured that g_s is actually a minimum of g and not a maximum. This condition is satisfied if

$$\frac{d^2}{dt^2}(g(t)) = 2a = (g_1 - 2g_2 + g_3) > 0.$$

The point x_s corresponding to t_s is given as follows:

$$x_s = x^n + t_s(x^n - x^0).$$

Now consider the position of x_s on the line joining x^n and x^0. Let $\mu d_p = (x^n - x^0)$ and $x_s = x^n + \lambda^p d_p$. Then by (2-5)

$$x^n + \lambda^p d_p = x^n \pm d_p \sqrt{2(f(x^n) - f(x_s))}$$

$$= x^0 + d_p \sqrt{2(f(x^0) - f(x_s))}.$$

The minus sign is to be used if x_s is between x^0 and x^n. Hence

$$(x^n - x^0) + d_p(\pm \sqrt{2(f(x^n) - f(x_s))} - \sqrt{2(f(x^0) - f(x_s))} = 0$$

$$d_p' A(x^n - x^0) + d_p' A d_p(\pm \sqrt{2(f(x^n) - f(x_s))} - \sqrt{2(f(x^0) - f(x_s))} = 0.$$

But $(x^n - x^0) = \mu d_p$, therefore,

$$\mu = \mp \sqrt{2(f(x^n) - f(x_s))} + \sqrt{2(f(x^0) - f(x_s))}.$$

Consider the case when x_s is between x^0 and x^n.

$$\mu = +\sqrt{2(f(x^n)-f(x_s))} + \sqrt{2(f(x^0)-f(x_s))}$$

$$= \sqrt{2(f(x^0)-f(x^n)+f(x^n)-f(x_s))} + \sqrt{2(f(x_n)-f(x_s))}.$$

Hence

$$|\mu| \geq |\sqrt{2(f(x^0)-f(x^n)}| > |\lambda^i|$$

which implies that;

$$|\lambda^i|/|\mu| < 1 \text{ for all } i.$$

But this last result violates one of the conditions that must be satisfied before the substitution of the direction (x^n-x^0) can be made. Therefore, when minimizing along the search direction (x^n-x^0) starting from x^n, if the minimum occurs at a point between x^n and x^0 the direction (x^n-x^0) is not used in the next search cycle. If Δ is defined to be the maximum decrease in the function over any of the search directions used, then the above conditions can be stated as follows:

If either $g_3 \geq g_1$ and/or $(g_1-2g_2+g_3)(g_1-g_2-\Delta)^2 \geq \frac{1}{2}\Delta(g_1-g_3)^2$ then the same search directions should be used again and x^n should be used as the new initial point. Otherwise x_s should be used as the new initial point and (x^n-x^0) should be substituted for that direction along which the function decreased the most during the last search cycle.

Unfortunately this new modification eliminates a useful property of the earlier method. The previous method

would, for a positive definite quadratic, generate n conjugate directions after n search cycles and thus the minimum would be achieved on the next cycle. The modification lacks this quality because of the manner in which new search directions are generated. There is no guarantee that a newly generated direction that is used to replace another one might not itself be eliminated later in the process. Thus the property of convergence after n+1 searches will most likely be lost. This, though, may not be as serious a problem as it seems for most functions to be dealt with will not be quadratic anyway.

A suggested criterion for convergence is to test whether the function has decreased in value significantly over a search cycle. But, as has been stated previously, for certain functions this could produce a premature halt to the procedure. Powell has suggested that the following criterion be used [13].

1. Continue the iterative process until the change in each variable over one cycle is less than one tenth the required accuracy. Let the resulting point in the last cycle be a.

2. Increase each variable by ten times the required accuracy and repeat step one, producing the point b.

3. Minimize along the line joining a and b to obtain the point c. Stop the process if the components of (a-c) and (b-c) are all less than one tenth the required accuracy.

4. Otherwise replace d_1^k by (a-c) and start step one again.

It appears as if this convergence criterion is a very strict one. Since this method requires a large number of functional evaluations, a convergence criterion which is too strict could cause the computer to do a great deal more work than is necessary. Of course, the nature of the problem will dictate the amount of accuracy required, which will ultimately affect the proper choice of convergence criterion. But in all cases there must be some sacrifice in accuracy made to avoid too many evaluations of the function.

The following method was designed in an attempt to alleviate another problem that arises in Powell's procedure. The requirements that must be satisfied before a new direction can be defined are much too demanding for problems involving a large number of variables. The result is that frequently one set of directions is used over and over again which, as is readily apparent, is similar to the alternating variable method discussed previously. It is therefore desirable to reduce these requirements if the main characteristics of the method can be maintained.

C. ZANGWILL'S METHOD

Zangwill proposed the following revision of Powell's method in hopes of increasing what might be a slow rate of convergence in the former method [6]. The procedure

involves the use of two different sets of search directions. The set c_i, $i = 1,\ldots,n$, is the co-ordinate directions, normalized so that $|c_i| = 1$. These directions will remain unchanged throughout the entire procedure. The other set d_i^j, $i = 1,\ldots,n$, where $|d_i^j| = 1$, contains n linearly independent directions. These last vectors are used much in the same manner as the search directions employed in Powell's method. Thus it is this set that will change after each search cycle. The process is begun from the initial point x_n^0. Then λ_n^0 is calculated to minimize $f(x_n^0 + \lambda_n^0 d_n^1)$ and x_{n+1}^0 is defined as the point $x_n^0 + \lambda_n^0 d_n^1$. The value of t initially is set equal to one. The procedure then becomes iterative. Thus for the first iteration t, the point x_{n+1}^0, and the directions d_i^1, $i = 1,\ldots,n$, are all known. In general, for the k^{th} iteration assume that t, the point x_{n+1}^{k-1}, and the directions d_i^k, $i = 1,\ldots,n$, are given. The k^{th} iteration proceeds as follows:

(i) Find α to minimize $f(x_{n+1}^{k-1} + \alpha c_i)$. Update t so that t is replaced by t+1, if $i \leq t \leq n$, and t is replaced by 1, if $t = n$. If $\alpha \neq 0$ let $x_0^k = x_{n+1}^{k-1} + \alpha c_i$. If $\alpha = 0$ repeat step (i). If step (i) is repeated n times in succession then no progress has been made when searching over the n co-ordinate directions. This will happen only when the minimum has been reached and thus the process should be halted. In this case the point at which the function is a minimum is x_{n+1}^{k-1}.

(ii) For $i = 1,\ldots,n$, calculate λ_i^k to minimize $f(x_{i-1}^k + \lambda_i^k d_i^k)$ and set $x_i^k = x_{i-1}^k + \lambda_i^k d_i^k$. Let $d_{n+1}^k = (x_n^k - x_{n+1}^{k-1})/||(x_n^k - x_{n+1}^{k-1})||$. Calculate λ_{n+1}^k to minimize $f(x_n^k + \lambda_{n+1}^k d_{n+1}^k)$ and set $x_{n+1}^k = x_n^k + \lambda_{n+1}^k d_{n+1}^k$. Set $d_i^{k+1} = d_{i+1}^k$ for $i = 1, \ldots,n$.

Now go to the k+1 iteration.

It can readily be seen that step (ii) is very similar to Powell's method which was discussed earlier.

Zangwill has proven the following important theorem concerning the above method.

THEOREM: Let f be a quadratic function with a positive definite Hessian A. The above procedure stops at an optimal point in step (i) of iteration k where $k \leq n$. (Recall that Powell's method could guarantee this type of convergence only if the linear independence of the search direction is maintained.)

PROOF: The proof will be by induction. Assume that at the beginning of the k^{th} iteration the method has generated k mutually conjugate directions, $d_{n-k+1}^k,\ldots,d_n^k$. The way in which this is done has been discussed earlier in this paper. Assume that the procedure does not stop during step (i). Therefore, a new point has been generated, which implies that $x_{n+1}^{k-1} \neq x_0^k$. Since $x_0^k = x_{n+1}^{k-1} + \alpha c_t$ where $\alpha \neq 0$ and was chosen to minimize $f(x_{n+1}^{k-1} + \alpha c_t)$ then $f(x_0^k) < f(x_{n+1}^{k-1})$. Also since x_n^k was generated by n minimizing searches beginning from x_0^k then $f(x_0^k) \geq f(x_n^k)$. Therefore, $f(x_n^k) \geq f(x_0^k) > f(x_{n+1}^{k-1})$ which implies that $x_n^k \neq x_{n+1}^{k-1}$. And thus $(x_{n+1}^{k-1}$

$- x_n^k) \neq 0$. During the k-1 iteration $d_{n-k+1}^k, \ldots, d_n^k$ were used as the last k search directions since $d_{i+1}^{k-1} = d_i^k$. Therefore the points x_{n+1}^{k-1} and x_n^k were found by minimizing in the k dimensional space spanned by $d_{n-k+1}^k, \ldots, d_n^k$. Therefore, as was shown previously, $d_{n+1}^k = x_{n+1}^{k-1} - x_n^k \neq 0$ is conjugate to the directions $d_{n-k+1}^k, \ldots, d_n^k$. Therefore by the n^{th} search cycle n conjugate search directions have been generated.

It remains to be shown that n conjugate vectors are linearly independent. Assume that they are not independent, which implies:

$$d_j^{k\prime} = \sum_{i \neq j} \alpha_i d_i^{k\prime}$$

$$d_j^{k\prime} A d_j^k = \sum_{i \neq j} \alpha_i d_i^{k\prime} A d_j^k = 0.$$

But since A was assumed to be positive definite, $d_j^{k\prime} A d_j^k > 0$, which is a contradiction. Thus the assumption that the vectors are linearly dependent must be false. Since the above argument holds for k = 1 the induction proof is complete.

Obviously since most functions of interest will not be quadratic away from the minimum this method will not generate true conjugate directions. For general functions it is not certain that this method is more efficient than the other methods presented earlier, when away from the minimum. However in the neighborhood of the minimum this latest method promises to be the most favorable thus far available, of those using only function evaluations.

IV. STEEPEST DESCENT AND NEWTON'S METHOD

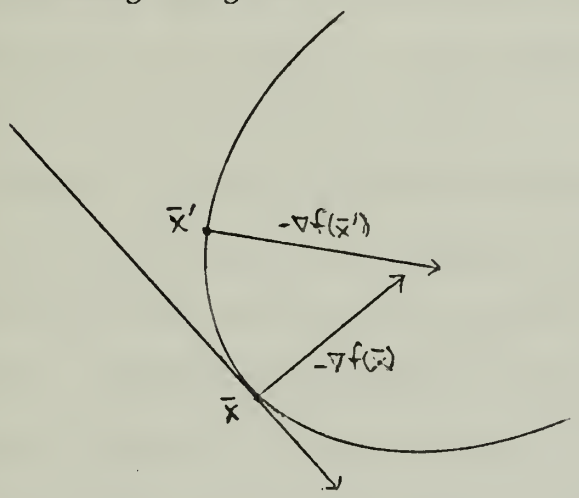
The method of Steepest Descent and Newton's method have been placed together in this chapter, though the theory and procedures of these methods are quite dissimilar. The Steepest Descent method uses only information about the first partial derivatives while Newton's method requires knowledge of the second partial derivatives.

They have been combined in this chapter because they are the two classical approaches to the minimization problem that remain useful today. They differ so much from the more recent methods that they deserve to be in a classification of their own. Many of the more recent methods have been devised as improvements of these two.

A. STEEPEST DESCENT BY CAUCHY

This method was one of the first techniques suggested to solve the problem under consideration. The main principle involved here is the use of $-\nabla f(x)$ as the search direction from x. This selection seems natural since a search in this direction from the point x insures that the function will at least initially decrease most rapidly. When far from the minimum this direction seems to be the most useful for it offers the opportunity to approach the minimum in one step rather than having to use n search directions as in the methods discussed previously. Unfortunately, as the minimum is approached this direction tends to be less and less useful.

One reason for this is that round off errors and inaccuracy in determining the gradient can have a great effect upon the search directions. This is demonstrated by the following diagram.



From the above diagram it can be seen that the direction that should be used and the one that is actually used might be almost perpendicular. Another problem for some functions is that this method may generate directions that cause the search to oscillate about the principle axis of the function with very little progress made in each search. Problems such as these significantly reduce the effectiveness of this method.

This method can be used in either of two ways, a fixed step size or by minimizing along the search direction. The latter technique requires minimization along $-\nabla f(x)$. The fixed step method sets the distance which is traveled along the search direction. The function is evaluated at this new point and this value is compared with the value of the function at the previous point. As long as the function is decreased the new point replaces

the previous one and the gradient is evaluated at this point to determine the next search direction. If one of these steps fails to reduce the value of the function then the step size is reduced and the process continued. Convergence is assumed to occur when the step size is reduced below a specified limit.

Both of these methods have their relative advantages. The fixed step technique requires fewer function evaluations while the minimization process should converge more rapidly. Unfortunately, though, neither method will, in general, proceed very rapidly when close to the minimum. In 1957, Booth suggested that the point nine tenths the distance to the minimum along the search direction should be used instead of the actual minimum. The purpose of this is to attempt to reduce the oscillation about the principle axis which is typical of this classical method. This simple procedure does reduce the problem but not enough to make the whole procedure useful for general functions [3].

B. NEWTON'S METHOD [3,9]

Sometimes the Hessian matrix may be known for the function to be minimized. Since, as has been suggested, it may be useful to take full advantage of all the information obtained about the function, it may be wise to search for a method which incorporates this second partial derivative information. Newton's method is just such a technique.

By using a second order Taylor expansion for a quadratic function the following equation is produced:

$$f(x) = f(\bar{x}+h) = f(\bar{x}) + \sum_{j=1}^{n} h_j \left[\frac{\partial f}{\partial x_j}\right]_{\bar{x}} + \tfrac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} h_j h_k \cdot \left[\frac{\partial^2 f}{\partial x_j \partial x_k}\right]_{\bar{x}}$$

where $\bar{x}$ is the point at which f is a minimum. But,

$$\frac{\partial f}{\partial x_i} = \left[\frac{\partial f}{\partial x_i}\right]_{\bar{x}} + \sum_{j=1}^{n} h_j \left[\frac{\partial^2 f}{\partial x_j \partial x_i}\right]_{\bar{x}} \quad i = 1,\ldots,n.$$

At the minimum $(\partial f/\partial x_1) = 0$ and therefore,

$$\frac{\partial f}{\partial x_i} = \sum_{j=1}^{n} h_j \left[\frac{\partial^2 f}{\partial x_j \partial x_i}\right]_{\bar{x}} \quad i = 1,\ldots,n. \qquad (3\text{-}1)$$

Let $\partial f/\partial x_i = g_i$ and $G_{jk} = \partial^2 f/\partial x_j \partial x_k$. Then equation (3-1) implies

$$g = Gh \text{ or } h = G^{-1}g.$$

Therefore

$$\bar{x} = x - G^{-1}g, \text{ since } x = \bar{x} + h.$$

The set of equations (3-1) must be solved to yield h. To do this the gradient at the current point must be known and the matrix of second partial derivatives must be available and evaluated at the minimum, $\bar{x}$. This last requirement poses some problems because, in general, the actual minimum must be known before this can be done. But if this point is known then there is no problem.

Newton used the preceding arguments to devise an iterative method to solve the minimization problem. When the search has reached the neighborhood of the minimum the matrix G is evaluated at the current point rather than the actual minimum. If in the neighborhood of the minimum the function approximates a quadratic, the matrix G tends toward a constant matrix and thus evaluating G at the current point should give a reasonable approximation to G when evaluated at the minimum.

Newton's method has been shown to be a very useful and powerful minimization technique. But like all techniques it does have its limitations. For example, progress toward the minimum is assured only if G is positive definite and the method may actually diverge for general functions. Another problem is the time required to generate G and G^{-1} and the storage space needed for these matrices. Since the matrix G is only used as an approximation, the time problem can be somewhat reduced. This can be done by calculating G and G^{-1} only after each k iterations rather than for each new step. Some of these problems are dealt with in the following method.

C. MODIFIED NEWTON'S METHOD [6]

The method now to be presented is a further refinement of the Newton's method that was just discussed. It was specifically designed to alleviate certain problems that the original Newton's method could not handle. One such

problem that will be dealt with is the occurrence of a Hessian matrix that might not be positive definite.

This method requires the selection of λ^i to minimize $f(x^i+\lambda^i d_i)$. As such this process is very similar to most of the methods discussed thus far; the key, though, is in the selection of d_i, the search direction. This is done according to the following rules:

1. If H_i, the current approximation to A^{-1}, has a negative eigenvalue then d_i should be chosen to satisfy the following:

$$d_i^! H_i d_i < 0 \text{ and } d_i^! \nabla f \leq 0. \tag{3-2}$$

2. If all the eigenvalues of H_i are nonnegative then choose d_i such that either,

$$H_i d_i = 0 \ , \quad d_i^! \nabla f < 0 \tag{3-3}$$

or

$$H_i d_i = -\nabla f. \tag{3-4}$$

Consider the first situation,

(i) $\partial f(x^i+\lambda^i d_i)/d\lambda_i = d_i^! \nabla f \leq 0$ at $\lambda^i = 0$

(ii) $\partial^2 f(x^i+\lambda^i d_i)/\partial\lambda_i^2 = d_i^! H_i d_i < 0$ at $\lambda^i = 0$.

Now (i) implies that f, at least initially, decreases in the search direction. If (ii) remains valid as $\lambda^i \to \infty$ then, obviously, the function is ever decreasing and thus must approach $-\infty$. But if this is the case then the minimum has been found. Otherwise there must be some place along d_i

where H_i becomes positive definite or semidefinite. Once this area is reached then rule number two can be applied until such a time as another area is reached where H_i has a negative eigenvalue.

Now consider the second situation. A similar argument holds. In this case, $d^2(f(x^i+\lambda^i d_i))/d\lambda^2 = d_i' H_i d_i = d_i' 0 = 0$, which also implies that unless an area of positive definiteness or semidefiniteness is reached along d_i the value of the function will again go to $-\infty$. Equation (3-3) obviously defines just the search direction given in the section of Newton's method.

In practice then this method employs the usual Newton search direction when in a region in which the Hessian is positive definite. In other regions the method selects new directions which should take the search process into an area where A is positive definite.

In general, therefore, this modification should improve the behavior of Newton's method when away from the minimum. It should also be able to solve a more general class of problems than the classical Newton's method. Unfortunately, however, it does not completely remove all the problems arising in the use of Newton's method. The most significant disadvantage of these latest two methods is that both require a great deal of information concerning the function. For some functions it could be just as time consuming to compute second partial derivatives as it is to solve the problem by some other procedure. Also

difficulties arise in problems of large dimension. Inverting an n x n matrix requires a great deal of work if it can be done at all. For these reasons other methods have been developed to approximate $H = A^{-1}$

V. VARIABLE-METRIC METHODS

This is the last major classification of methods to be discussed in this paper. As such it represents some of the most recent developments in the field of function minimization. The theory behind these methods is rather simple in concept. It involves making better and better approximations to the matrix $H = A^{-1}$, where the function to be minimized is assumed to be approximated by a quadratic function f of the form: $f(x) = x'Ax/2 + b'x + c$. If f were actually quadratic then knowledge of A^{-1} would allow the minimum to be reached in one step, as was shown by equation (1-1), Chapter I. Thus any method which can generate this matrix would indeed be valuable.

A. DAVIDON, FLETCHER, AND POWELL [4,8]

The original work in this area was presented by Davidon in 1959, but Fletcher and Powell took Davidon's original method and improved upon it to the extent that theirs has become one of the more popular and reliable methods available for minimization. Though most of the original work was done by Davidon, the notation and arguments by Fletcher and Powell are more concise and will be used in the discussion to follow.

Let $g_i = \nabla f(x^i)$ and $d_i = -H_i g_i$ where H_i is the i^{th} approximation to A^{-1}. The matrix A is assumed to be positive definite and H_o, the first estimate of H, is usually

selected as the identity matrix. Note that this selection for H_0 produces an initial search direction that is simply that of the Steepest Descent method.

Let us define the vectors

$$p_i = \lambda^i d_i = x^{i+1} - x^i$$

and

$$q_i = g_{i+1} - g_i.$$

The vector p_i is the step to the minimum along d_i from the point x^i, and q_i is the corresponding change in the gradient.

For this method it is desired to repeatedly update the matrix H_i to make better and better approximations to A^{-1}. Consider a recursion formula which generates H_i's with the following properties. The set of vectors $p_0,\ldots,p_k$, are linearly independent and they are eigenvectors of $H_{k+1}A$ with one as eigenvalues. Then, obviously, H_nA will have n linearly independent eigenvectors with eigenvalue one. But this can occur only if $H_nA = I$ and thus $H_n = A^{-1}$ as desired. It will be established that the following recursion formula satisfies these requirements.

$$H_{i+1} = H_i + \frac{p_i p_i'}{p_i' q_i} + \frac{-(H_i q_i)(H_i q_i)'}{q_i' H_i q_i} \qquad (4\text{-}1)$$

First let us show that p_i is an eigenvector of $H_{i+1}A$ with eigenvalue one. To do this consider:

$$q_i = g_{i+1} - g_i$$

$$= (Ax^{i+1} + b) - (Ax^i + b)$$

$$= Ax^{i+1} - Ax^{i}$$

$$= Ap_i .$$

Hence

$$H_{i+1}Ap_i = H_{i+1}q_i$$

$$= H_i q_i + \frac{p_i p_i' q_i}{p_i q_i} - \frac{(H_i q_i)(q_i' H_i q_i)}{(q_i' H_i q_i)} \text{ , by (4-1),}$$

$$= H_i q_i + p_i - H_i q_i$$

$$= p_i . \qquad (4\text{-}2)$$

Finally, if the following two results can be established then the desired properties of p_i, $i = 1,\ldots,n$, will be established.

$$p_i' Ap_j = 0 \quad 0 \leq i < j < k \qquad (4\text{-}3)$$

$$H_k Ap_i = p_i \quad 0 \leq i < k. \qquad (4\text{-}4)$$

Equation (4-3) implies conjugacy which has been shown to require linear independence, and equation (4-4) is the desired result concerning eigenvalues and eigenvectors. Equations (4-3) and (4-4) will be established by induction. Consider equation (4-4) with $k = 1$,

$$H_1 Ap_0 = p_0$$

by (4-2). Consider (4-3) with $k = 2$,

$$p_0' Ap_1 = (p_1' Ap_0)' = (-\lambda^1 g_1' H_1 Ap_0)'$$

or

$$p_0' Ap_1 = -\lambda^1 g_1' p_0 = 0, \qquad (4\text{-}6)$$

since we minimize along p .

Now assume that

$$p_i'Ap_j = 0, \quad 0 \leq i < j < k,$$

and

$$H_iAp_i = p_i, \quad 0 \leq i < k. \tag{4-7}$$

Consider:

$$\begin{aligned} g_k &= Ax^k + b \\ &= A(x^{i+1}+p_{i+1}+\ldots+p_{k-1}) + b \\ &= g_{i+1} + A(p_{i+1}+\ldots+p_{k-1}). \end{aligned}$$

Hence, we see that

$$\begin{aligned} p_i'g_k &= p_i'g_{i+1} + p_i'Ap_{i+1}+\ldots+p_i'Ap_{k-1} \\ &= p_i'g_{i+1}, \quad \text{by (4-6)}, \\ &= 0. \end{aligned}$$

But,

$$\begin{aligned} 0 = p_i'g_k &= (H_kAp_i)'g_k \\ &= p_i'AH_kg_k \\ &= p_i'A(-d_k) \\ &= -p_i'A(p_k/\lambda^k); \end{aligned}$$

and hence

$$p_i'Ap_k = 0, \quad 0 \leq i < k.$$

But this is equivalent to the following:

$$p_i'Ap_j = 0 \ , \ 0 \leq i < j < k+1. \tag{4-8}$$

Also,

$$H_{k+1}Ap_i = H_kAp_i + \frac{p_kp_k'Ap_i}{p_i'q_k} - \frac{(H_kq_k)(q_k'H_kAp_i)}{q_k'H_kq_k}$$

$$= p_i + \frac{p_kp_k'Ap_i}{p_i'q_k} - \frac{(H_kq_k)(q_k'H_kAp_i)}{q_k'H_kq_k}$$

$$0 \leq i < k$$

or by (4-8),

$$H_{k+1}Ap_i = p_i \ , \ 0 \leq i < k. \tag{4-9}$$

By equations (4-5), (4-6), (4-8), and (4-9) the induction proof has been completed. Thus $H_n = A^{-1}$, as was desired.

An obvious question at this point is what motivated the choice of the recursion formula? Consider the following:

Let $d_1,\ldots,d_n$ be mutually conjugate directions. Then

$$(\sum_{i=1}^{n} \beta_i d_i d_i')Ad_s = \beta_s d_s d_s'Ad_s = d_s \quad \text{if } \beta_s = \frac{1}{d_s'Ad_s}.$$

Therefore,

$$A^{-1} = \sum_{i=1}^{n} \frac{d_i d_i'}{d_i'Ad_i}$$

$$= \sum_{i=1}^{n} (\lambda^i)^2 d_i d_i' / (\lambda^i)^2 d_i'Ad_i$$

$$= \sum_{i=1}^{n} p_i p_i'/p_i'Ap_i$$

$$= \sum_{i=1}^{n} p_i p_i'/p_i'q_i$$

Thus we see that the second term in the recursion formula was selected to make the approximation approach A^{-1}.

It will now be shown that the third term on the right side of equation (4-1) is added as a correction factor. As was shown previously it was necessary that $H_{i+1}Ap_i = p_i$ to make this method valid. Consider the following:

$$H_{i+1} = H_i + p_i p_i'/p_i'q_i + C_i.$$

It will now be determined what form C_i must take in order to satisfy the condition that $H_{i+1}Ap_i = p_i$.

$$p_i = H_{i+1}Ap_i$$

$$= H_iAp_i + p_i p_i'Ap_i/p_i'q_i + C_iAp_i$$

$$= H_iAp_i + p_i + C_iAp_i.$$

Hence,

$$H_iAp_i = -C_iAp_i$$

or

$$H_iq_i = -C_iq_i.$$

A solution for C_i to this equation is:

$$C_i = -H_iq_iz'/z'q_i$$

where z is an arbitrary vector that is not perpendicular to q_i.

For this choice of C_i:

$$C_i q_i = \frac{-H_i q_i z' q_i}{z' q_i} = -H_i q_i \text{ , as desired.}$$

But since it is desired that C_i be symmetric, z is set equal to $H_i q_i$.

Therefore,

$$C_i = -H_i q_i (q_i' H_i)/(q_i' H_i q_i), \text{ which is as desired.}$$

Now consider the search direction d_i.

$$-d_i' g_i = g_i' H_i g_i.$$

If it can be shown that $g_i' H_i g_i$ is positive then it is evident that the search direction is always in the direction of decreasing function values and thus the λ^i's can be chosen positive. But if H_i can be shown to be positive definite then $g_i' H_i g_i > 0$ as desired. Since H_0 is chosen to be positive definite it remains to be shown by induction argument that by (4-1) if H_i is positive definite then so is H_{i+1}. Consider

$$x' H_{i+1} x = x' H_i x + \frac{(x' p_i)(p_i' x)}{p_i' q_i} - \frac{(x' H_i q_i)(q_i' H_i x)}{q_i' H_i q_i}$$

$$= \frac{(x' H_i x)(q_i' H_i q_i) - (x' H_i q_i)(q_i' H_i x)}{q_i' H_i q_i}$$

$$+ \frac{(x' p_i)^2}{p_i' q_i} .$$

But $(x'H_i x)(q_i'H_i q_i) \geq (x'H_i q_i)(q_i'H_i x)$, by Schwartz's inequality. Therefore

$$x'H_{i+1}x \geq (x'q_i)^2/p_i'q_i,$$

with equality only if x and q_i are parallel. Obviously, $(x'q_i)^2$ is greater than or equal to zero so it remains to be shown that $p_i'q_i > 0$ for their quotient to be greater than zero.

$$\begin{aligned} p_i'q_i &= p_i'(g_{i+1} - g_i) \\ &= -p_i'g_i, \text{ by minimization along } d_i, \\ &= -\lambda^i d_i'g_i \\ &= \lambda^i g_i'H_i g_i > 0, \end{aligned}$$

since H_i was assumed to be positive definite.
Hence, $x'H_{i+1}x > 0$ for all nontrivial x; this implies H_{i+1} is positive definite, and thus the proof is complete.

By proving that H_i is positive definite for all i, it has been assured that for a quadratic function that this method is completely stable and will produce the minimum in at most n steps. The last part of this section concerning the positive definiteness was specifically due to Fletcher and Powell.

The ease with which this method can be applied and its strong stability make it one of the most useful methods thus far discussed. As with all methods it will only be approximate for functions which are more general than the quadratic

function discussed above; but it would be natural to assume that it still would usually out perform the other techniques.

Of course, there are some problems present in this method also. As with most, Davidon's technique requires determining the minimum along a given direction. While this may or may not be a major problem it will require additional function evaluations which must be considered. Also there could be a storage problem, especially for large dimensional problems, since at each step an n x n matrix must be saved. It is the first of these problems that the next method attempts to avoid. It is not unusual for the necessary function and gradient evaluations to use half of the total computer time required for the solution of the problem. Thus, if the number of evaluations is reduced, without altering the basic method itself, it would be expected that the result would be a more efficient method.

B. MURTAGH AND SARGENT

This method, devised by Murtagh and Sargent [10], uses a recursion formula similar to that employed by Davidon, Fletcher, and Powell. The recursion formula to be is used in generating A^{-1} is

$$H_{k+1} = H_k + (p_k - Hq_k)(p_k - Hq_k)'/q_k'(p_k - Hq_k). \quad (4\text{-}10)$$

As will be shown, the advantage of this method is that it does not require minimization along each search direction. This new formula can be developed as follows.

Let us assume as before that the function is approximated by a quadratic, $f(x) = x'Ax/2 + b'x + c$. Then,

$$g(x_k) = Ax_k + b$$

$$g(x_{k-1}) = Ax_{k-1} + b$$

and

$$g(x_k - x_{k-1}) = A(x_k - x_{k-1})$$

or

$$g_k = Ap_k$$

where

$$g_k = g(x_k) - g(x_{k-1})$$

and

$$p_k = x_k - x_{k-1},$$

as defined earlier.

Now let H be an approximation to A^{-1}. If H were exact then $Hq_k = p_k$. But since H is not exact there is an error involved. Let e be this error, $e = p_k - Hq_k$, and consider now adding ΔH to H so that $(H+\Delta H)q_k = p_k$. Then,

$$Hq_k + \Delta Hq_k = p_k$$

$$\Delta Hq_k = p_k - \Delta Hq_k = e.$$

If ΔH is chosen so that each column is a multiple of e then ΔHq_k is also a multiple of e. Since it is desirable for H to remain symmetric the following choice for ΔH is made:

$$\Delta H = mee'$$

in which m is a constant to be determined. We require $mee'q_k = e$. Hence,

$$m = 1/e'q_k = 1/(p_k - Hq_k)'q_k$$

and

$$\Delta H = (p_k - Hq_k)(p_k - Hq_k)'/(p_k - Hq_k)'q_k.$$

This produces the following recursion formula for H.

$$H_{k+1} = H_k + (p_k - Hq_k)(p_k - Hq_k)'/q_k'(p_k - Hq_k). \quad (4\text{-}11)$$

It should be recalled at this time that for a quadratic function the step to the minimum is given as follows:

$$\bar{x} - x_k = -A^{-1}g(x_k). \quad (4\text{-}12)$$

A similar search direction will be employed in this method with the addition of an arbitrary scalar, α_k, so that:

$$p_k = -\alpha_{k-1}H_{k-1}g_{k-1}. \quad (4\text{-}13)$$

The scalar is added to this formula because the step in (4-12) may at times provide a poor estimate of the distance to the minimum.

The advantage of the recursion formula developed in (4-11) is that it is not required to minimize the function along each search direction. As was stated previously, this requirement for minimization was a disadvantage of Davidon's method. Unfortunately, though, this alteration does reduce the stability. Murtagh and Sargent prove the

following theorem concerning the convergence of their method [10].

THEOREM: Assume that the function $f(x)$ is defined on $U \subset E^n$ and is such that

i. $f(x)$ is continuous on $\Omega = \{x | x \in U;\ f(x) \leq c\}$ and Ω is closed and bounded.

ii. $f(x)$ has continuous second derivatives on $\Omega' = \{x | x \in U;\ f(x) < c\}$ and there is a Λ such that $||H(x)|| \leq \Lambda$, $x \in \Omega'$.

Starting at any point $x_0 \in \Omega'$ with $g(x_0) \neq 0$, we generate a sequence $x_0, x_1, \ldots, x_k, x_{k+1}, \ldots$ from

$$p_{k+1} = x_{k+1} - x_k = -\alpha_k H_k g_k.$$

Then if the matrices H_k satisfy the conditions:

$$\rho ||g_k|| \leq ||H_k g_k|| \leq \sigma ||g_k||$$

$$|g_k' H_k g_k| \geq \delta ||g_k||\ ||H_k g_k||$$

where ρ, σ, and δ are fixed positive constants, it is always possible to choose a finite nonzero α_k at each step such that:

$$f(x_k) - f(x_{k+1}) \geq \varepsilon \alpha_k g_k' H_k g_k > 0$$

with ε a fixed positive constant less than unity. With α_k so chosen, the sequence (x_k) lies in Ω' and tends to $\Omega^* = \{x | x \in \Omega';\ g(x) = 0\}$ in the sense that the distance $d(x_k, \Omega^*)$ of x_k from Ω^* tends to zero as $k \to \infty$.

Murtagh and Sargent's method satisfies the conditions of the above theorem if H_k is positive definite for all k. A theorem by Caratheodory (1967) can be used to establish conditions under which H_k will be positive definite.

Let

$$z_k = p_k - H_{k-1}q_k \text{ and } c_k = q_k'z_k.$$

Define

$$H(t) = H_{k-1} + tz_kz_k'/c_k . \tag{4-14}$$

By Caratheodory's theorem H(t) is positive definite in the range $0 \leq t \leq 1$ if H_{k-1} is positive definite and H(t) is nonsingular over this range. Since $H_k = H_{k-1} + z_kz_k'/c_k$, to show that H_k is positive definite by assuming that H_{k-1} is positive definite it is sufficient to show that H(t) is nonsingular for $0 \leq t \leq 1$. From (4-14),

$$\det H(t) = \det H_{k-1}\,(1 + tz_k'H_{k-1}^{-1}z_k/c_k)$$

$$= \det H_{k-1}\,(1 + tz_k'H_{k-1}^{-1}(p_k - H_{k-1}q_k)/c_k)$$

$$= \det H_{k-1}\,(1 + tz_k'H_{k-1}^{-1}p_k/c_k - tz_k'q_k/c_k)$$

$$= \det H_{k-1}\,\left(1 - t + \frac{(tz_k'H_{k-1}^{-1}(-\alpha_{k-1}H_{k-1}g_{k-1}))}{c_k}\right)$$

$$= \det H_{k-1}\,(1 - t - \alpha_{k-1}tz_k'g_{k-1}/c_k).$$

It is necessary that

$$(1 - t - \alpha_{k-1}tz_k'g_{k-1}/c_k) > 0 \quad 0 \leq t \leq 1 \tag{4-15}$$

for H_k to be positive definite and nonsingular. Since α_{k-1} is positive; it is necessary that

$$z_k' g_{k-1}/c_k < 0. \tag{4-16}$$

Since a positive definite matrix (usually I) can be chosen for H_0, equation (4-16) is a necessary condition for H_k to be a positive definite matrix for all k. Consider the numerator and denominator of (4-16) separately.

$$\begin{aligned} z_k' g_{k-1} &= (p_k - H_{k-1} q_k)' g_{k-1} \\ &= (-\alpha_{k-1} H_{k-1} g_{k-1} - H_{k-1} q_k)' g_{k-1} \\ &= (-\alpha_{k-1} g_{k-1}' H_{k-1} - (g_k' - g_{k-1}') H_{k-1}) g_{k-1} \\ &= (1-\alpha_{k-1}) g_{k-1}' H_{k-1} g_{k-1} - g_k' H_{k-1} g_{k-1} \end{aligned} \tag{4-17}$$

$$\begin{aligned} c_k &= z_k' q_k \\ &= z_k' (g_k - g_{k-1}) \\ &= (-\alpha_{k-1} g_{k-1}' H_{k-1} - g_{k-1}' H_{k-1} + g_{k-1}' H_{k-1}) \cdot \\ &\qquad g_k - z_k' g_{k-1} \\ &= -\alpha_{k-1} g_{k-1}' H_{k-1} g_{k-1} - g_k' H_{k-1} g_k + g_{k-1}' H_{k-1} g_k \\ &\qquad - z_k' g_{k-1} \\ &= (1-\alpha_{k-1}) g_{k-1}' H_{k-1} g_k - g_k' H_{k-1} g_k - z_k' g_{k-1} \end{aligned} \tag{4-18}$$

Solving for α_{k-1} in equation (4-17) and substituting this into (4-18) produces the following result

$$g_{k-1}'H_{k-1}g_{k-1}c_k = z_k'g_{k-1}(g_{k-1}'H_{k-1}g_k - g_{k-1}'H_{k-1}g_{k-1})$$

$$-(g_k'H_{k-1}g_kg_{k-1}'H_{k-1}g_{k-1} - (g_{k-1}'H_{k-1}g_k)^2). \quad (4\text{-}19)$$

Since H_{k-1} is positive definite, $g_{k-1}'H_{k-1}g_{k-1} > 0$, and the sign of the quantity on the left side of equation (4-19) depends upon the sign of c_k. Schwartz's inequality shows that:

$$-(g_k'H_{k-1}g_kg_{k-1}'H_{k-1}g_{k-1} - (g_{k-1}'H_{k-1}g_k)^2) \geq 0. \quad (4\text{-}20)$$

Now assume that $z_k'g_{k-1} > 0$ and examine (4-17),

$$z_k'g_{k-1} = (1-\alpha_{k-1})g_{k-1}'H_{k-1}g_{k-1} - g_k'H_{k-1}g_{k-1} > 0.$$

If $0 < \alpha_{k-1} < 1$,

$$(1-\alpha_{k-1})g_{k-1}'H_{k-1}g_{k-1} > g_k'H_{k-1}g_{k-1}$$

and hence

$$g_{k-1}'H_{k-1}g_{k-1} > g_k'H_{k-1}g_{k-1}.$$

If $\alpha_{k-1} = 1$,

$$-g_k'H_{k-1}g_{k-1} > 0$$

whence

$$g_k'H_{k-1}g_{k-1} < 0.$$

But H_{k-1} is positive definite and therefore $g_{k-1}'H_{k-1}g_{k-1} > 0$. Therefore,

$$g_{k-1}'H_{k-1}g_{k-1} > g_k'H_{k-1}g_{k-1}.$$

If $\alpha_{k-1} > 1$,

$$(1-\alpha_{k-1}) < 0.$$

Hence,

$$(1-\alpha_{k-1})g_{k-1}'H_{k-1}g_{k-1} > g_k'H_{k-1}g_{k-1}$$

which implies that,

$$g_k'H_{k-1}g_{k-1} < 0.$$

Again,

$$g_{k-1}'H_{k-1}g_{k-1} > 0$$

which implies,

$$g_{k-1}'H_{k-1}g_{k-1} > g_k'H_{k-1}g_{k-1}.$$

Therefore $z_k'g_k > 0$ implies that $g_{k-1}'H_{k-1}g_{k-1} > g_k'H_{k-1}g_{k-1}$, which by (4-15) implies,

$$z_k'g_{k-1}(g_{k-1}'H_{k-1}g_k - g_{k-1}'H_{k-1}g_{k-1}) < 0. \qquad (4.21)$$

By (4-20) and (4-21) then $c_k < 0$. Therefore if $z_k'g_{k-1} > 0$, then $z_k'g_{k-1}/c_k < 0$ which is as required by (4-16). Similarly $c_k > 0$ implies that $z_k'g_{k-1} < 0$, which once again satisfies condition (4-16). Unfortunately, though, $z_k'g_{k-1} \leq 0$ and $c_k < 0$ can occur simultaneously. Then condition (4-16) is violated and H_k will not be positive definite. Since it is necessary that this property is maintained this can cause some serious problems. Consider the results

of taking a step with $\alpha_{k-1} = 1$. Equations (4-17) and (4-18) imply:

$$z_k' g_{k-1} = -g_k' H_{k-1} g_{k-1}$$

$$c_k = -g_k' H_{k-1} g_k - z_k' g_{k-1}.$$

If $z_k' g_{k-1}$ is positive then, obviously, c_k is negative and the positive definite requirement is satisfied. Thus it might be wise first to take this step with $\alpha_{k-1} = 1$ and to test $z_k' g_{k-1} > 0$. If this is the case then there is no problem and H_k should be updated by the given recursion formula. If this step is taken but $z_k' g_{k-1} \leq 0$ then generally the function has decreased. This in itself is desirable since the process has reached a "better" point. If $z_k' g_{k-1} < 0$ then it should be tested for $c_k > 0$. If this holds then condition (4-16) again has been satisfied and thus the recursion formula should be applied. Here, though, it is wise to add a test to ensure that c_k is not too close to zero, which would contribute additional problems. If this becomes the case or $z_k' g_{k-1}/c_k \geq 0$ then it is necessary to start again with a new H_0 from the latest best point. Murtagh and Sargent suggest two possible choices for this new H_0, either I or the previous H_i. The first choice, of course, is simply starting over again with no information about H. This could be useful if H_k has begun to accumulate misinformation concerning the function. But, in general, it would probably be best not to destroy all the previous information.

Murtagh and Sargent offer a number of algorithms employing the methods they devised [10]. The one that they found to be the most useful involved the checks discussed thus far and, in addition, certain checks to ensure that the conditions of their theorem were met. By the criterion of fewest function evaluations required this last method was generally found to be the most efficient. Again this is as has been anticipated because the need to minimize along the search directions was reduced. It was found that the conditions of the theorem of Murtagh and Sargent was far less restrictive then requiring actual minimization.

If function evaluations was the only criterion then it could be said that generally Murtagh and Sargent's method was superior to Davidon, Powell, and Fletcher's. But because of all the tests that must be made the method must surely be more difficult to program and, outside of function evaluations, more time consuming to run. In addition, since at times the conditions for positive definiteness can fail and a new H_0 must be selected, the convergence of the method will obviously be slowed down. Should this resetting of H be required too often there is no doubt that all advantages this method might have would be lost.

C. PEARSON'S CLASS OF VARIABLE METRIC METHODS

In this section will be presented a class of related methods devised by Pearson [11]. Included in this class is

the method invented by Davidon. Though the recursion formulas of these methods are different, their development is closely related as will be shown in the following.

In the previous section it was shown that for a quadratic function $q_k = Ap_k$ where $q_k = g_k - g_{k-1}$ and $p_k = x_k - x_{k-1}$. Therefore, $Hq_k = p_k$, where $H = A^{-1}$. Consider the following possibility. Assume $H_j q_i = p_i$, for $i = 1,\ldots,j$, where H_j is the j^{th} approximation to A^{-1}. If H_j can be updated so that $H_n q_i = p_i$, for $i = 1,\ldots,n$, and if the set $\{p_i;\ i = 1,\ldots,n\}$ is linearly independent then $H_n = A^{-1}$. This can be seen from the following:

Assume,

$$H_n q_i = p_i,$$

for $i = 1,\ldots,n$. But,

$$q_i = Ap_i,$$

for $i = 1,\ldots,n$ therefore,

$$(H_n A)p_i = p_i,$$

for $i = 1,\ldots,n$ and hence

$$(H_n A - I)p_i = 0.$$

But the set $\{p_i\}$, $i = 1,\ldots,n$, is linearly independent, which implies that $H_n = A^{-1}$. Now define the search directions as before, $d_i = H_i' g_i$.

Then,

$$q_s' d_i = q_s' H_i' g_i, \text{ for } s = 1,\ldots,i-1$$

$$= p_s' g_i.$$

Thus, if $p_s'g_i = 0$, then

$$q_s'd_i = 0 \qquad (4\text{-}21)$$

for s = 1,...,i-1. But consider what occurs when the function is minimized along each search direction.

$$H_n q_i = p_i$$

$$p_j'AH_n q_i = p_j'Ap_i,$$

for j = 1,...,i-1. But,

$$p_j'AH_n q_i = p_j'q_i$$

$$= p_j'(g_i - g_{i-1})$$

$$= p_j'g_i - p_j'g_{i-1}, \quad \text{for } j = 1,\ldots,i-1.$$

If j = i-1, then by minimization along d_{i-1},

$$p_{i-1}'g_{i-1} = 0$$

and thus

$$p_{i-1}'(g_i - g_{i-1}) = p_{i-1}'g_i.$$

If j < i-1 then,

$$p_j'(g_i - g_{i-1}) = p_j'g_i - p_j'g_{i-1}.$$

In either case,

$$p_s'g_i = 0 \text{ for } s = 1,\ldots,i-1$$

implies

$$p_j'Ap_i = p_j'g_i - p_j'g_{i-1} = 0 \text{ for } j = 1,\ldots,i-1,$$

and therefore each new direction generated is conjugate to the previous ones. Thus if the condition (4-22) is satisfied

the method generates conjugate directions. The problem then is to find solutions to $H_k q_i = p_i$, for $i = 1,\ldots,k-1$. To do this define the following matrices:

$$Q_i = (q_1 \cdots q_{i-1})$$

$$P_i = (p_1 \cdots p_{i-1}).$$

In this notation the problem then is to find solutions to

$$H_i Q_i = P_i, \tag{4-23}$$

for $i = 1,\ldots,n$. Consider,

$$H_i = P_i(Q_i'MQ_i)^{-1}Q_i'M + H_0(I - Q_i(Q_i'M^*Q_i)^{-1}Q_i'M^*) \tag{4-24}$$

$$H_iQ_i = P_i(Q_i'MQ_i)^{-1}(Q_i'MQ_i) + H_0IQ_i - H_0Q_i(Q_i'M^*Q_i)^{-1}$$

$$(Q_i'M^*Q_i)$$

$$= P_i$$

where M and M* are arbitrary. Thus (4-24) defines a solution to (4-23). It was given that M and M* were arbitrary but this is not completely true since $Q_i'MQ_i \neq 0$ and $Q_i'M^*Q_i \neq 0$. Obviously, if M and M* are chosen to be positive definite then $Q_i'MQ_i$ and $Q_i'M^*Q_i$ are unequal to zero. There are two matrices that seem to be likely choices for M and M*. First there is H_0 which, of course, can and will be selected to be positive definite; and then there is A^{-1} which is assumed to be positive definite.

Since nothing in (4-24) specifies otherwise, M and M* can be chosen independently. By doing so, four different

forms of (4-24) can be produced. Pearson makes use of a lemma called the Bordered Inverse Lemma to produce the following recursion formulas from equation (4-24).

$$H_{i+1} = H_i + (p_i - H_i q_i)(p_i')/p_i' q_i \qquad (4\text{-}25)$$

$$H_{i+1} = H_i + (p_i - H_i q_i)(H_i q_i)'/q_i' H_i q_i \qquad (4\text{-}26)$$

$$H_{i+1} = H_i + p_i' p_i/p_i' q_i - (H_i q_i)(H_i q_i)'/q_i' H_i q_i. \qquad (4\text{-}27)$$

Notice that (4-27) is exactly (4-1) which was Fletcher, Powell and Davidon's recursion formula.

It can readily be seen that equations (4-25) and (4-26) both produce H_i's that will not be symmetric. This is, of course, a slight disadvantage since it will require additional storage space, as compared with Davidon's method which produces symmetric matrices. This, though, should only be significant in problems involving a large number of variables. In general, the results so far indicate that the three recursion formulas given above produce similar results. For some functions it may be necessary to replace the current approximation of A^{-1} with the positive definite matrix that was originally chosen. This happens if the approximation becomes singular as the minimum is approached. Generally, though, this is not required.

VI. CONCLUSIONS

In the writing of this paper the author has studied the literature. Of course, it has by no means covered every possible method by which the unconstrained minimization problem may be solved. But an attempt has been made to offer as wide a coverage as possible of the different techniques which are available. The methods included are those which have been found to be the most reliable in solving actual problems. Research with computers on specific problems have shown that no one method is guaranteed to out perform all others on every problem. In general then, the greatest difficulty might be the actual selection of the method to be employed.

To make this decision it is important to consider all the information about the function which is available. This includes such things as having second partial derivatives which may be computed, or the knowledge of only the gradients, or only having access to the function values. Generally it has been found that gradient methods are usually the most reliable, but this is in reference to methods applied to functions for which the gradient is available from analytic expressions. On occasion, though, for some functions the gradient is only calcuable through numerical methods. When this is the case the accuracy of the entire method is greatly reduced. In fact, in these cases it is best, as a rule, to use one of the methods employing only function values.

For most functions there may be a number of methods that can be used to obtain the minimum. Thus if all that is required is the proper minimum then the problem of selection may be greatly reduced. But in practice there are, of course, other important factors which must be considered.

Computer time for the solution is one of these vital factors. Consider, for example, the Steepest Descent method. For certain functions this method may have an extremely slow rate of convergence. But if this technique leads to the correct minimum then it must be included among the methods from which the one method to be used is selected. But to select this method in such a case would be a serious mistake. There may be another method which could solve the prbblem in one tenth the time. With all other factors equal this other method would obviously be the better choice.

In general, though, the relative advantages of each method are unknown for any given function. It would be best to study the function to be minimized before any selection of a technique is made. If any special characteristic of the function can be identified then it may be possible to make a wiser selection of the method to be used.

The author has found the book by Box [3] and the book by Kowalik and Osborne [9] to be particularly helpful.

APPENDIX A: LINEAR SEARCH TECHNIQUES

In this appendix will be presented a number of methods for finding the minimum of a function along a given line. Since many of the minimization techniques discussed in this paper require one of these methods their importance cannot be minimized.

1. FIBONACCI SEARCH [3,9]

Assume that we decide to make N function evaluations within the interval (x_1,x_2) where a minimum is known to exist. Assume the original neighborhood is designated (x_1^1,x_2^1). Then define the following points:

$$x_3^1 = \frac{F_{N-2}}{F_N}(x_2^1-x_1^1) + x_1^1$$

$$x_4^1 = \frac{F_{N-1}}{F_N}(x_2^1-x_1^1) + x_1^1,$$

where F_n is a Fibonacci number, defined by the following relations:

$$F_0 = F_1 = 1$$

$$F_n = F_{n-1} + F_{n-2} \quad n \geq 2.$$

If $f(x_3^1) > f(x_4^1)$ then the minimum must lie between x_3^1 and x_2^1. Otherwise the minimum must lie between x_4^1 and x_1^1. Consider $(x_4^1-x_1^1)$ and $(x_2^1-x_3^1)$:

$$x_4^1-x_1^1 = \frac{F_{N-1}}{F_N}(x_2^1-x_1^1)$$

$$x_2^1 - x_3^1 = x_2^1 - \frac{F_{N-2}}{F_N}(x_2^1 - x_1^1) - x_1^1 \qquad \text{(A-1)}$$

$$= (1 - F_{N-2}/F_N)x_2^1 - (1 - F_{N-2}/F_N)x_1^1$$

$$= \frac{(F_N - F_{N-2})x_2^1}{F_N} - \frac{(F_N - F_{N-2})x_1^1}{F_N} \quad .$$

Thus,

$$x_2^1 - x_3^1 = \frac{F_{N-1}}{F_N}(x_2^1 - x_1^1) \quad . \qquad \text{(A-2)}$$

Thus (A-1) and (A-2) show that, regardless of which interval the minimum has been restricted to, the length of the interval is (F_{N-1}/F_N) times the length of the original interval. The end points of the interval containing the minimum are the relabeled, x_1^2 and x_2^2. The process is then repeated using the following general formulas.

$$x_3^i = \frac{F_{N-1-i}}{F_{N+1-i}}(x_2^i - x_1^i) + x_1^i$$

$$x_4^i = \frac{F_{N-i}}{F_{N+1-i}}(x_2^i - x_1^i) + x_1^i$$

for i = 1,...,N-1. The final two points would, by the above formulas, coincide and thus should be offset by some small ε. The length of the final iterval to which the minimum is restricted is:

$$(x_2^1 - x_1^1)/F_N + \varepsilon.$$

Thus the accuracy to which the minimum is found depends upon the size of the original interval and the number of function evaluations to be made. By the nature of Fibonacci numbers it is only necessary to make one function evaluation per iteration after the initial iteration.

2. DAVIES, SWANN AND CAMPEY'S SEARCH TECHNIQUE [3,9]

In this method the function is approximated along the line by a quadratic. If the three points used to locate the minimum are separated by an interval greater than some specified size, the operation is repeated with a smaller interval.

For this method an initial step size is decided upon depending upon the estimated distance to the minimum from the current point. This step size should be about one fourth this estimated distance. The initial step is taken toward the minimum and the function is evaluated at this new point. If the function has increased then cut the step size and begin again from the initial point. This is done until a point is found for which the function has decreased. The step size is then doubled and a new step is taken from the latest point. This process is continued until a function increase is located. At this time the current step size is cut in half and a step is taken from the last point at which the function decreased. The last four points found are thus equally spaced say, s units apart, and define an interval in which the minimum must lie. The end point of this interval furtherest from the point at which the function has the smallest value is discarded and

the remaining three points are used to approximate the minimum. Let x_1, x_2, and x_3 be these three points with f_1, f_2, and f_3 be the respective function values at these points. Let us assume a quadratic approximation for f on the line,

$$f(t) = at^2 + bt + c,$$

in which the values -1, 0, 1 for t correspond to values at x_1, x_2, and x_3 respectively.

Thus,

$$a = (f_1 - 2f_2 + f_3)/2$$

$$b = (f_1 - f_3)/2$$

$$c = f_2.$$

Therefore,

$$f(t) = \frac{(f_1 - 2f_2 - f_3)t^2}{2} + \frac{(f_3 - f_1)\ t}{2} + f_2.$$

The minimum of f(t) is at $t_s = (f_1 - f_3)/2(f_1 - 2f_2 + f_3)$. The point x_s corresponding t_s is $x_s = x_2 + t_s(x_2 - x_1)$. The function is then evaluated at this new point and the process is begun again with a reduced step size from the point at which the function had the smallest value. The process is continued until the change in successive approximations to the minimum is less than half the desired accuracy.

3. POWELL'S ALGORITHM [3,9]

This method differs from the previous method only in the manner of selecting the three points from which to interpolate the minimum. Assuming x_1 is the initial point

then $x_2 = x_1 + S$ is selected where S is a fixed displacement. The third point, x_3, is chosen as follows:

$$x_3 = x_1 + 2S \text{ if } f_1 \geq f_2$$

$$x_3 = x_1 - S \text{ if } f_1 < f_2.$$

Using these three points another quadratic interpolation is performed as outlined above. If the new point differs from the point where the function had its smallest value by less than the required accuracy then this point is assumed to be the desired minimum. Otherwise the point at which the function is the largest is discarded and a new interpolation is made using the remaining three points.

4. DAVIDON'S CUBIC INTERPOLATION

Davidon uses a cubic interpolation based on values for the function and its gradient at two points of the line [4, 13].

Assume that the value of the function and its gradient are known at two points, x and y, where $y = x + \alpha d$. This method calls for using a cubic interpolation as follows. Let,

$$f(t) = at^3 + bt^2 + ct + d \qquad 0 \leq t \leq \alpha$$

where

$$f(x) = f(0) = d$$

$$f(y) = f(\alpha) = a\alpha^3 + b\alpha^2 + c\alpha + d$$

$$\nabla f(x)'s = g_{sx} = \left(\frac{df}{dt}\right)_0 = c$$

$$\nabla f(y)'s = g_{sy} = \left(\frac{df}{dt}\right)_\alpha = 3a\alpha^2 + 2b\alpha + c.$$

Therefore,

$$a = (g_{sy}\alpha + g_{sx}\alpha - 2(f_y - f_x))/\alpha^3$$

$$b = (3(f_y - f_x) - \alpha g_{sy} - 2g_{sx}\alpha)/\alpha^2$$

$$c = g_{sx}$$

$$d = f_x.$$

Thus,

$$\frac{df}{dt}\text{'s} = g_{st} = 3at^2 + 2bt + c$$

$$= 3((g_{sy}\alpha + g_{sx}\alpha - 2(f_y - f_x)t^2/\alpha^3$$

$$+2t(3(f_y - f_x) - \alpha g_{sy} - 2g_s\alpha)/\alpha^2 + g_{sx}$$

$$= g_{sx} - 2t(g_{sx} + B)/\alpha + t^2(g_{sx} + g_{sy} + 2B)/\alpha^2$$

where $B = 3(f_x - f_y)/\alpha + g_{sx} + g_{sy}$.

From the quadratic formula it is found that the desired zero is at

$$t_m = \frac{+\frac{2}{\alpha}(g_{sx} + B) + \sqrt{\frac{4}{\alpha}(g_{sx}^2 + B^2 + 2g_{sx}B) - \frac{4}{\alpha^2}(g_{sx}^2 + g_{sx}g_{sy} + 2Bg_{sx}}}{\frac{2}{\alpha^2}(g_{sx} + g_{sy} + 2B)}$$

$$= \frac{\alpha(g_{sx} + B + Q)}{(g_{sx} + g_{sy} + 2B)}$$

where

$$Q = \sqrt{B^2 - g_{sx}g_{sy}}$$

$$= \alpha \left[\frac{(g_{sx}+g_{sy}+2B) - (g_{sy}+B-Q)}{g_{sx}+g_{sy}+2B} \right]$$

$$= \alpha \left[1-\frac{(g_{sy}+B-Q}{(g_{sx}+g_{sy}+2B)} \right].$$

And thus,

$$t_m = \alpha \left[1-\frac{(g_{sy}+Q-B)}{(g_{sy}-g_{sx}+2Q)} \right] . \qquad \text{(A-3)}$$

Since the condition for a minimum along S is that the component of the gradient along S is zero, the above equation (A-3) gives an estimate for this minimum. Thus the estimate for k such that f(x + kS) is a minimum is:

$$k = \alpha \left[1- \frac{(g_{sy}+Q-B)}{(g_{sy}-g_{sx}+2Q)} \right].$$

For the solution to this problem then it is only necessary to ensure that a reasonable choice is made for α.

BIBLIOGRAPHY

1. Balokrishnan, A. V., Symposium on Optimization, 1969, Springer-Verlag, 1970.

2. Box, M. J., "A Comparison of Several Current Optimization Methods, and the Use of Transformations in Constrained Problems," The Computer Journal, v. 9, p. 67-77, January 1966.

3. Box, M. J., Nonlinear Optimization Techniques, Oliver and Boyd, 1969.

4. Argonne National Laboratory, Variable Metric Method for Minimization, by William C. Davidon, May 1959.

5. Davidon, William C., "Variance Algorithm for Minimization," The Computer Journal, v. 10, p. 406-410, February 1968.

6. Fiacco, A. V., and McCormick, G. P., Nonlinear Programming: Sequential Unconstrained Minimization Techniques, p. 156-178, Wiley, 1968.

7. Fletcher, R., Optimization, Academic Press, 1969.

8. Fletcher, R., and Powell, M. J. D., "A Rapidly Convergent Descent Method for Minimization," The Computer Journal, v. 6, p. 163-168, July 1963.

9. Kowalik, J., and Osborne, M. R., Methods for Unconstrained Optimization Problems, American Elsevier, 1968.

10. Murtagh, B. A., and Sargent, R. W. H., "Computational Experience with Quadratically Convergent Minimization Methods," The Computer Journal, v. 13, p. 185-194, May 1970.

11. Pearson, J. D., "Variable Metric Methods of Minimization," The Computer Journal, v. 12, p. 171-179, May 1969.

12. Powell, M. J. D., "An Efficient Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives," The Computer Journal, v. 7, p. 155-162, May 1964.

13. Powell, M. J. D., "An Iterative Method for Finding Stationary Values of a Function of Several Variables," The Computer Journal, v. 5, p. 147-151, July 1962.

14. Rosenbrock, H. H., "An Automatic Method for Finding the Greatest or Least Value of a Function," The Computer Journal, v. 3, p. 175-184, October 1960.

15. Wilde, D. J., Optimum Seeking Methods, Prentice-Hall, 1964.

16. Zangwill, W. I., Minimizing a Function Without Calculating Derivatives," The Computer Journal, v. 10, p. 293-299, November 1967.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3.	Professor Frank Faulkner, Code 53Fa Department of Mathematics Naval Postgraduate School Monterey, California 93940	1
4.	ENS Gary C. Meyer, USN 14415 Frankton Avenue Hacienda Heights, California 91745	1

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)

Naval Postgraduate School
Monterey, California 93940

2a. REPORT SECURITY CLASSIFICATION

UNCLASSIFIED

2b. GROUP

3 REPORT TITLE

THE UNCONSTRAINED MINIMIZATION PROBLEM

4 DESCRIPTIVE NOTES (Type of report and, inclusive dates)

Master's Thesis; June 1971

5. AUTHOR(S) (First name, middle initial, last name)

Gary C. Meyer

6 REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
June 1971	82	16

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Naval Postgraduate School Monterey, California 93940

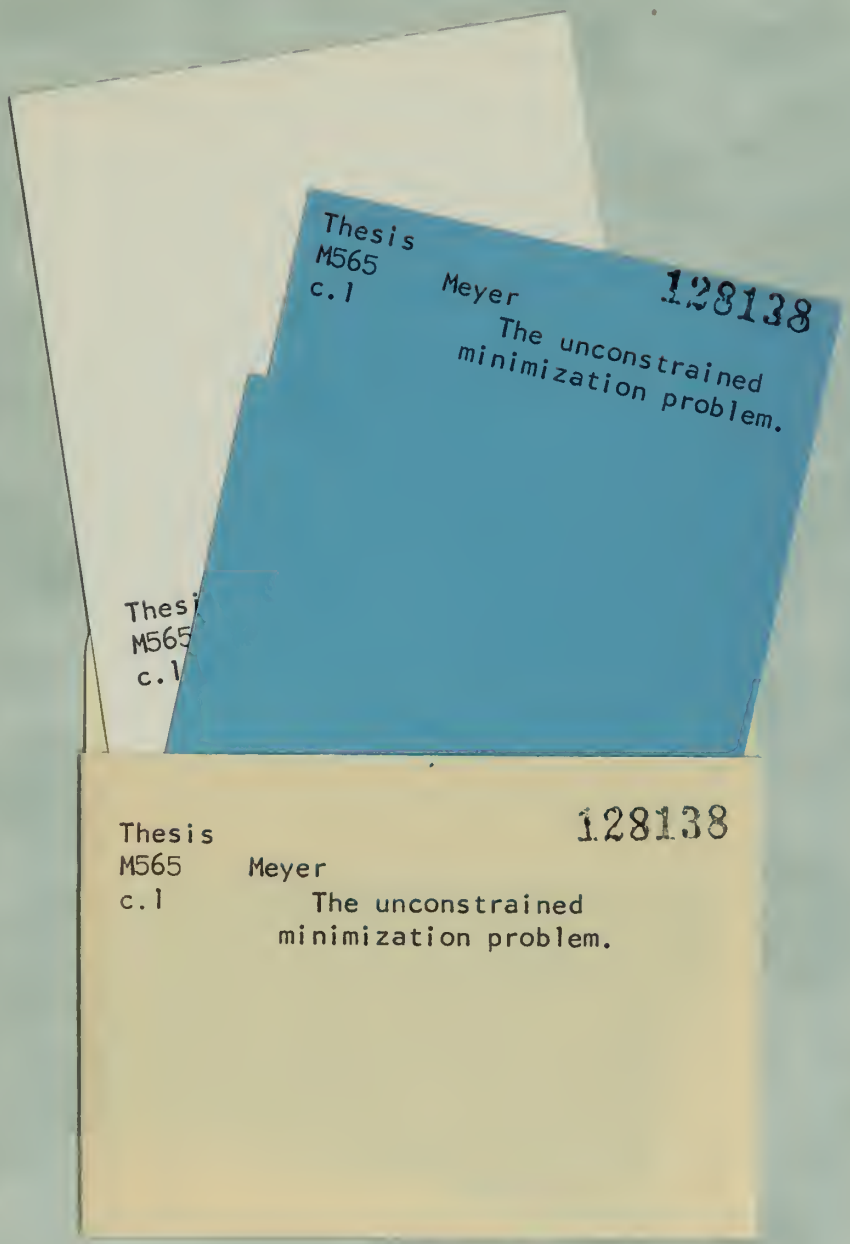
13. ABSTRACT

Considered within this paper is the problem of minimization of a function of unconstrained variables. A wide variety of solutions to this problem is presented and the possible advantages of each method are discussed. For the purpose of this paper these techniques are divided into four broad categories: general search directions; conjugate search directions; Cauchy's Steepest Descent and Newton's method; and variable metric methods.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Minimization						
Unconstrained Variables						
Functional Analysis						


Thesis
M565
c.1
Meyer
128138
The unconstrained
minimization problem.
Thesis
M565
c.1
128138
Meyer
The unconstrained
minimization problem.